UNLIMITED

(2)

RSRE
MEMORANDUM No. 4336

ROYAL SIGNALS & RADAR ESTABLISHMENT

FURTHER EXPERIMENTS IN VARIABLE FRAME RATE ANALYSIS FOR SPEECH RECOGNITION

Authors: S M Peeling & K M Ponting

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

UNLIMITED

CONDITIONS OF RELEASE

0067559 BR-113377

********** DRIC U

********** DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

********** DRIC N

DCAF CODE 090996

Royal Signals and Radar Establishment

Memorandum 4336

Further Experiments in Variable Frame Rate Analysis for Speech Recognition

S.M. Peeling and K.M. Ponting

February 21, 1990

Abstract

The application of a simple variable frame rate analysis to the RSRE Airborne Reconnaissance Mission system, a continuous speech recognition system based on phone-level hidden Markov models, is described. Results are presented which show that performance using the variable frame rate technique and triphone models can be better than that obtained using triphone models and full frame rate data. The variable frame rate technique requires considerably less processing time.

Contents

List of Figures

List of Tables

1 Introduction

This memo describes the application of a simple VFR analysis to the RSRE Airborne Reconnaissance Mission (ARM) system. This is a continuous speech recognition system based on phone-level hidden Markov models (HMMs) which has been developed at the RSRE Speech Research Unit.

In a companion paper ([9]) it was shown that variable frame rate (VFR) analysis can be used to actually improve the recognition performance of certain hidden Markov models (HMMs). For completeness the general descriptions from [9] are replicated in Sections 2, 3 and 4.

This memo concentrates on experiments which used VFR analysis on triphone models, described in more detail in Section 5.

The results quoted here come from three speakers, unlike those in [9] which came from a single speaker.

2 The Variable Frame Rate Algorithm

This section will briefly describe the nature of the data, what VFR analysis is, and its application to automatic speech recognition.

Assume that at any "instant" in time the speech signal can be represented by an ordered set of numbers, or feature vector. This "instant" is assumed to be short enough that the properties of the speech signal do not change significantly. Any utterance, or collection of words, can then be described as a succession of feature vectors (sometimes referred to as frames). There are areas where the speech signal is relatively constant and hence successive feature vectors will be very similar. In other areas the signal may change rapidly and hence successive feature vectors will be different.

In order to reduce the processing time, one obvious solution is to reduce the data (frame) rate. However, parts of the signal which are changing rapidly contain valuable information and so need to be retained. For this reason it is necessary to employ some method of data reduction which actually depends on the data. Variable frame rate coding is such a technique.

One of the first uses of VFR for data reduction in automatic speech recognition is described in [1]. In that paper the authors describe several different VFR algorithms. This memo has utilised the simplest of those.

The VFR algorithm has been designed to retain all the input feature vectors when they are changing most rapidly and omit a high proportion when they are relatively constant. A subset of the feature vectors is selected, thus avoiding the need for deciding how to combine vectors. All that is required is the calculation of some measure of similarity between two feature vectors and the comparison of this similarity measure with a threshold. The

most common similarity measure used is the Euclidean distance which is used in all the experiments quoted here.

In the simple version of the algorithm the distance is computed between the last retained feature vector and the vector under consideration. The current vector is then omitted if the distance is less than the threshold. With this approach, a threshold less than the minimum distance (zero) results in vectors being retained at the original frame rate. A threshold set to the maximum distance (effectively infinity) would result in a single vector being output, and an intermediate threshold provides a variable frame rate dependent on the speech data.

Specifically, if $D(i,j)$ is the distance between the previously selected frame j and the current frame i, and the threshold is T, then the rule is to select frame i as the next output frame if :-

$$D(i,j) \geq T$$

In some applications different thresholds are used which decrease with time so the likelihood of outputting a frame increases with time. This application has used a single threshold but has set an upper bound of 50 (referred to as the duplication factor) on the number of frames which can be represented by any one output frame, thus effectively incorporating a time constraint. This limit is only reached in long periods of silence and is necessary to ensure that they are not completely reduced.

3 Speech Representations

The speech data used were obtained by passing digitised speech signals through a 27 channel filter bank analyser at 100 frames per second. The filters are spaced on a non-linear frequency scale based on that in [4].

As specified in [10] all the HMMs used data which had the speech spectra replaced by a cepstral representation. The data used in this report consisted of 16 mel frequency cosine coefficients (MFCCs) together with an overall amplitude feature.

For the results reported here, the count of frames represented by a given feature vector after VFR analysis is appended to the vector as an additional feature. This results in 18 features for 16 MFCCs. The use of this additional feature provides a crude duration model, in that the model mean counts will reflect the average number of frames in the original analysis which are condensed to a single frame corresponding to the model states. The inclusion of this extra feature was shown in [9] to be beneficial.

4 Recognition And Scoring

The recognition algorithm used is a sub-word model implementation of a one-pass dynamic programming algorithm ([2]).

Whole ARM reports are processed including silences between sentences.

Scoring is based on a dynamic programming alignment at the phoneme level, taking account of the known sentence end times, with subsequent marking of words according to whether their constituent phonemes correctly line up (*cf* [8], [3]).

Recognition results are reported for two levels of syntactic constraint. All the phone results come from employing the *simple* syntax in which any sequence of triphones can be recognised. The word results are obtained from the *word* syntax which allows recognition of any sequence of non-speech sounds and words from the ARM vocabulary.

As with the results quoted in [9], these results are presented in terms of *% words correct* and *% word accuracy*. These are computed as follows, using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\%\ words\ correct = \frac{N - S - D}{N} \times 100,\ \%\ word\ accuracy = \frac{N - S - D - I}{N} \times 100$$

where N is the number of words in the test set, and S, D and I are the number of words substituted (i.e. recognised as the incorrect word), deleted and inserted respectively. The more interesting results are those in the columns headed "word accuracy" since these reflect more closely the level of performance which would be perceived by a user of the system.

As in [10] the training and test data were distinct sets of ARM reports. Unless otherwise stated, all the recognition results are for a 540 word test set.

5 HMMs And Triphone Models

The ARM system will not be described in full here. Further details can be found in [10], [11].

The theory and use of sub-word hidden Markov models for automatic speech recognition is now well established (eg [6]). These systems typically have distinct models corresponding to each phoneme in the language, which are combined according to a pronunciation dictionary to give whole word models for recognition. A large set of models is usually used, allowing different models for a given phoneme according to its immediate phoneme context (so-called triphone models).

The version of the ARM system described in the earlier paper ([9]) used a smaller set of models: four models for non-speech sounds; six models of short common words[1] and sixty-one models of the phonemes in the ARM dictionary (some phonemes have two distinct models, for syllable–initial and syllable–final consonants, which is the only context sensitivity embodied in that model set).

[1] of, or, in, at, air, oh (used instead of zero sometimes)

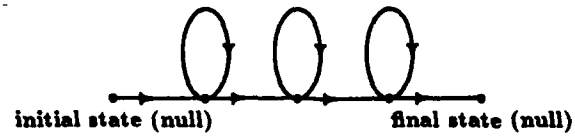

Figure 1: Topology of 3-state phone-level HMMs used in the ARM system

For the experiments reported here, the same set of function word and non-speech models is used, but full triphone models are used in place of the phoneme models. (There are approximately 1500 word-internal triphones in the ARM vocabulary; word boundary triphones are not used.)

In the earlier paper ([9]), it was shown that using three states per phoneme with the VFR analysis produced results nearly as good as the best obtained using the more expensive duration sensitive model topology (the "TI" topology). Therefore the work reported here uses a standard topology with three states per phoneme [2], and no skip transitions as shown in Figure 1.

All state output probability density functions of HMMs in the system are Gaussian with a diagonal (co)variance matrix. The same variance is used for all states of all models in this version of the system (the so-called Grand Variance) in order to reduce the total number of parameters to be estimated.

Initial estimates of HMM parameters were obtained from a small quantity of speech which had been hand labelled at the phoneme level. Standard HMM algorithms were then used to train context insensitive phoneme models on the full training set of 36 ARM reports (224 sentences, 1985 words), using much coarser labelling. [3] These context insensitive models are then used to provide initial values for reestimation of the corresponding triphone models as described in [12].

6 Results

Full results are quoted for two (male) speakers RKM and MJR. Some of the experiments were also repeated for the (female) speaker SRJ used in [9].

6.1 Effect Of Different Thresholds On Data Files

When applying the VFR technique to speech data it is useful to know what sort of data reduction is being obtained. In the ARM system training and testing files are dealt

[2] All models have three states except for the non-speech models which have only a single state; and the models for the function words "at", "in" and "of" which have six states.

[3] The data for the two male speakers was automatically labelled in breath groups. However for SRJ this labelling had been done at the sentence level, combining breath groups - it is not yet known how this difference affects performance.

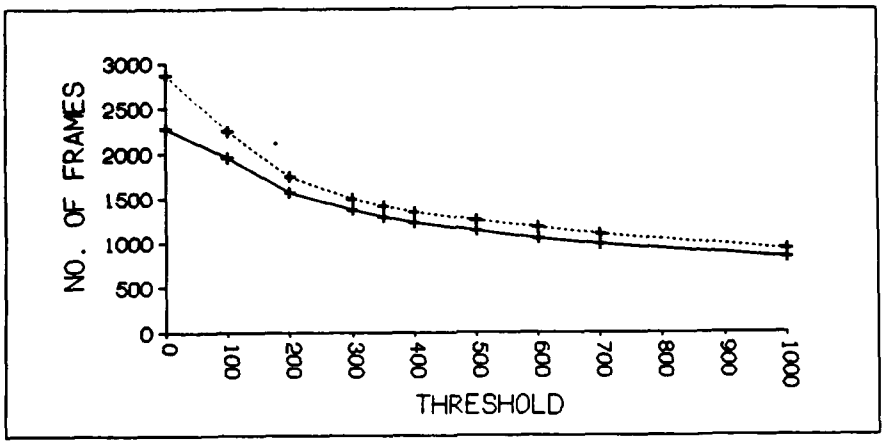

Figure 2: The effect of different thresholds on numbers of frames, from typical files, processed during training (solid line) and during testing (dotted line) for speaker RKM.

with differently. For training purposes, only the actual speech in the file is used - any silence, between labelled phrases or sentences, is ignored. During recognition however whole reports are processed, including silences (and breath noises etc) between sentences, unlike many other systems. The data reduction obtained, from using the VFR technique on data from the speaker SRJ, was shown in [9]. Similar reductions were obtained for the other two speakers. Figure 2 shows the effect of the threshold on the number of frames processed for two typical files for the speaker RKM.

In this figure, the solid line shows the amount of speech used in a typical training file. The dotted line shows the total amount of speech (including silences) used in a typical testing file. The more rapid data reduction at low thresholds for the testing file is due to the silences being discarded. Notice that even a relatively small threshold (about 300-400) is capable of almost halving the amount of speech data to be processed.

As stated in Section 5 the triphone models contained one, three or six states. During Baum-Welch reestimation for a particular utterance, a concatenated model is constructed from the models of the constituent triphones (and function words). Given the simple model topologies used, at least one frame of data is required for each state in the concatenated model, otherwise reestimation fails because there are no valid paths aligning the full sequence of states to input frames. Clearly, as the VFR threshold was increased the number of frames in a particular utterance decreased. This resulted in two problems. Firstly, even at quite low thresholds some of the utterances were too short to be modelled and hence could not be used for training purposes. However, relatively few utterances were involved (even at high thresholds) so there were still sufficient for training purposes.

The other problem arose when some of the hand labelled instances of phonemes used to "seed" the models were too short. Difficulties only arose at a threshold of 1000 when there were no valid hand labelled examples of several phonemes. This problem was overcome by locating and hand labelling a few longer examples of these phonemes which were then used in training at all thresholds. For speakers RKM and MJR about six triphones and two function words were involved. Because of the different annotation for speaker SRJ this problem did not arise.

6.2 Effect Of Different Thresholds On Processing Times

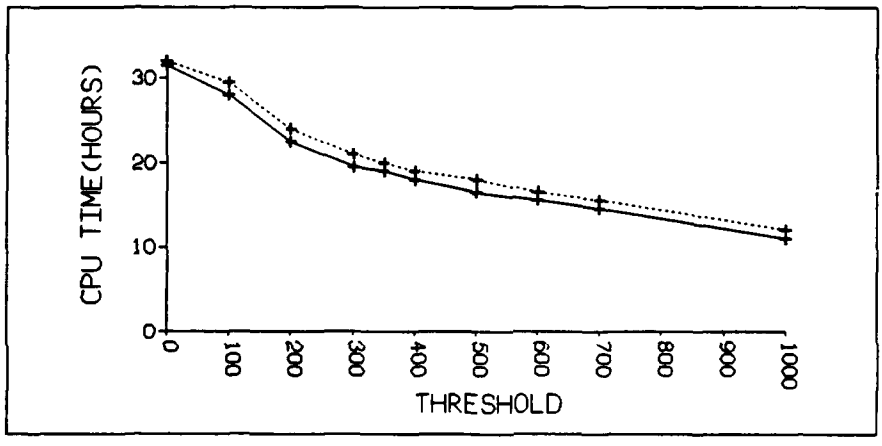

Figure 3: The effect of different thresholds on the processing time used in the training phase for speakers RKM (solid line) and MJR (dotted line).

Figure 3 shows how the processing time of the training phase decreases with increasing threshold. As expected, the shape of these lines are very similar to those in Figure 2. [4]

Even a threshold of 300 produces a significant reduction in computing time.

6.3 Triphone Models And Different VFR Thresholds

The recognition results for triphone models and different thresholds are shown in Table 1. Word accuracy results are shown in Figure 4 for the speakers RKM and MJR. It can be seen that using VFR techniques again gives an improvement in performance, but this increase is not as marked as in [9].

[4] The training time is proportional to the sum, over all utterances, of the number of states times the frame length of the utterance. The numbers of states do not change but the utterance lengths do.

VFR Threshold	Speaker	Phone		Word	
		correct	accuracy	correct	accuracy
0	RKM	54.1%	11.2%	92.0%	79.8%
100	RKM	54.8%	15.1%	92.8%	80.9%
200	RKM	55.7%	24.2%	93.0%	83.3%
300	RKM	55.5%	29.5%	93.1%	85.4%
350	RKM	55.7%	32.7%	92.8%	85.4%
400	RKM	54.1%	31.9%	92.6%	85.0%
500	RKM	54.1%	36.5%	90.2%	80.2%
600	RKM	51.3%	36.0%	87.6%	75.2%
700	RKM	49.1%	36.8%	83.0%	66.3%
1000	RKM	37.8%	31.3%	68.3%	35.0%
0	MJR	57.9%	16.0%	94.3%	83.0%
100	MJR	58.8%	20.9%	93.9%	83.9%
200	MJR	60.0%	32.3%	94.3%	86.7%
300	MJR	60.3%	36.0%	93.7%	88.0%
350	MJR	60.0%	37.8%	93.1%	86.5%
400	MJR	59.3%	38.6%	93.0%	87.6%
500	MJR	58.8%	41.8%	91.3%	83.9%
600	MJR	56.4%	41.7%	88.5%	78.3%
700	MJR	52.4%	40.6%	87.0%	72.2%
1000	MJR	42.8%	36.8%	72.6%	43.5%
0	SRJ	58.7%	23.5%	95.9%	86.5%
350	SRJ	59.5%	37.9%	95.7%	88.7%

Table 1: Recognition results for speakers and thresholds as shown.

Figure 5 shows that the improved performance was mainly due to there being fewer insertions. [5] It can be seen that the number of substitutions and deletions is virtually constant for VFR thresholds below 400. However there is a steady decrease in the number of insertions over the same range.

As the threshold increases above 400, substitutions and deletions begin to increase, as do the insertions.

[5] This is only shown for speaker RKM since the corresponding graphs for MJR were virtually identical.

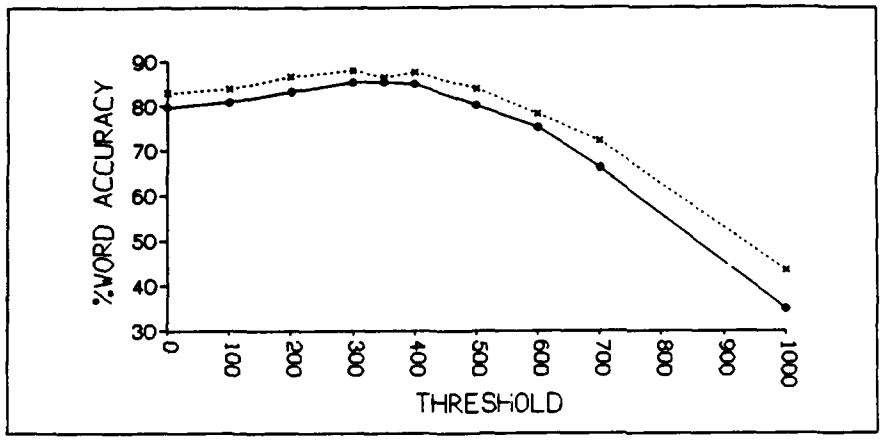

Figure 4: Word accuracy results for speakers RKM (solid line) and MJR (dotted line).

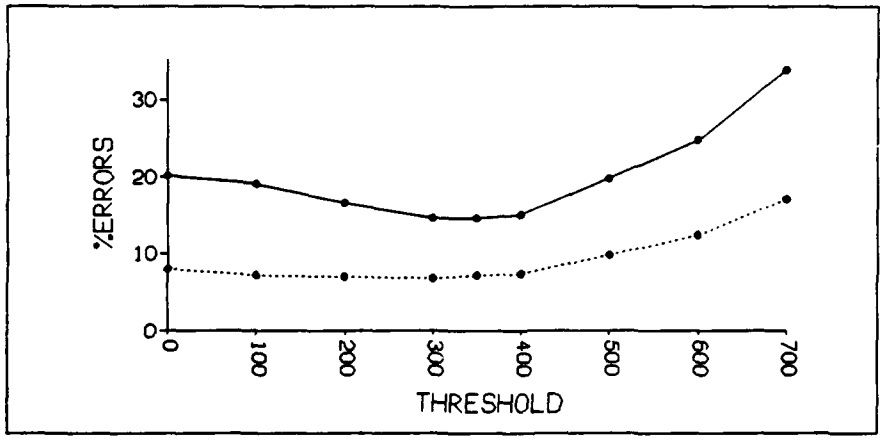

Figure 5: Percentage errors in word accuracy results for speaker RKM counting substitutions plus deletions (dotted line) and substitutions plus deletions plus insertions (solid line).

6.4 The Effect Of The DC Offset

The speech signals used have a large DC offset. Since one of the filters used in the filter bank analyser is centred at zero frequency, this offset feeds directly into the output of the analyser. Hence, when the MFCC coefficients are calculated they are all affected by this offset. This is not a very desirable state of affairs so experiments were conducted on the effect of removing this offset by omitting that channel from the analyser. These were only conducted for speakers RKM and MJR at two thresholds. The results are shown in Table 2.

VFR Threshold	Speaker	DC Offset	Phone		Word	
			correct	accuracy	correct	accuracy
0	RKM	yes	54.1%	11.2%	92.0%	79.8%
		no	55.3%	11.7%	92.4%	79.6%
350	RKM	yes	55.7%	32.7%	92.8%	85.4%
		no	56.9%	36.2%	92.8%	84.6%
0	MJR	yes	57.9%	16.0%	94.3%	83.0%
		no	57.9%	18.5%	92.4%	79.1%
350	MJR	yes	60.0%	37.8%	93.1%	86.5%
		no	60.6%	39.8%	92.4%	85.4%

Table 2: Recognition results for speakers and thresholds as shown, with and without DC offset.

From the results in Table 2 it can be seen that the only significantly different word accuracy result from removing the DC offset is obtained for speaker MJR at a threshold of zero, when the word accuracy decreases. In order to investigate this behaviour the means and standard deviations of the MFCC values were studied. For a typical test file for speaker MJR the means and standard deviations were calculated for 16 MFCCs with and without the DC offset. These values were calculated over the true speech in the test file, i.e. there were no silences, glitches or breath noises present, and also over the true silences in the test file. Graphs were then drawn of the means and standard deviations for the true speech and true silence, without the DC offset (Figure 6) and with the DC offset(Figure 7).

In both these figures, the error bars span a distance of twice the standard deviation. The mean value is at the centre of the bar. From these figures it can be seen that without the DC offset there is quite a lot of overlap between the values for true speech and true silence. The standard deviations of the MFCC values for true silence are considerably smaller than those for true speech when the DC offset is present. The effect of the DC offset appears to be to reduce both the standard deviations of the silence MFCCs and the overlap with the values for true speech. It is hoped to carry out further investigation into this behaviour in the future.

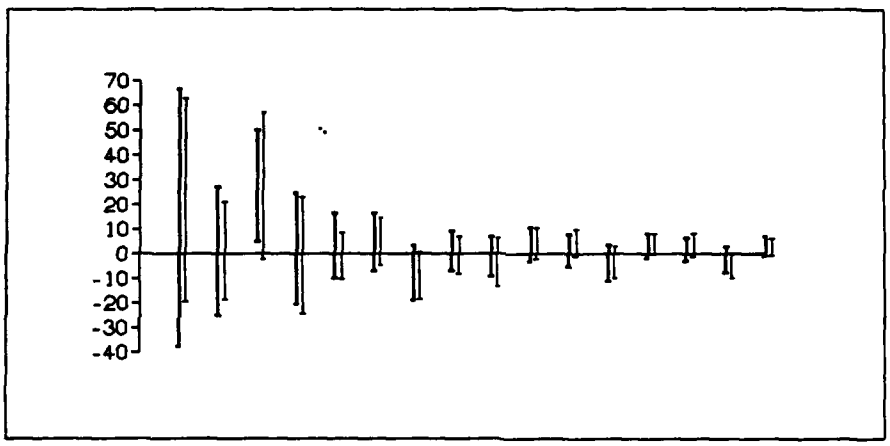

Figure 6: Means and standard deviations of MFCC values for 16 MFCC coefficients for a test file for speaker MJR at a threshold of zero without DC offset. The bold lines show the values for the true speech and the normal lines the values for true silence.

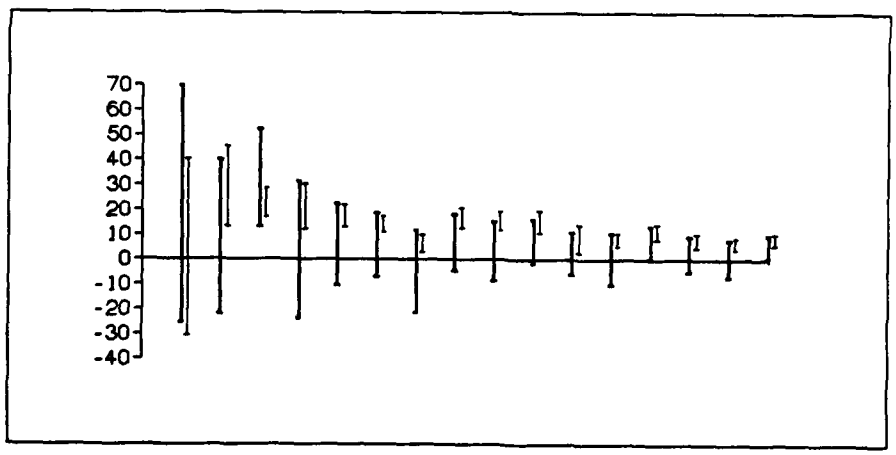

Figure 7: Means and standard deviations of MFCC values for 16 MFCC coefficients for a test file for speaker MJR at a threshold of zero and with DC offset. The bold lines show the values for the true speech and the normal lines the values for true silence.

6.5 Variance Weighting As An Alternative To VFR

It has been suggested, notably in Section 7.9 in [5], that VFR is capable of achieving such good performance because it downweights steady state regions, which typically occur in long vowels. In these steady state regions the effect of a relatively minor spectral difference is amplified by the number of frames over which it is repeated. Conversely, very rapid formant transitions at vowel-consonant boundaries may be crucial in identifying a consonant although only occupying a single frame. When trying to "match" an unknown word with a template, undue notice may be taken of the steady state regions, i.e. the vowels are matched more closely than the consonants. Using VFR analysis counterbalances this effect.

It has been suggested ([7]) that the effect of VFR in downweighting steady state regions could be emulated for non-VFR models and data by modifying the variances of the models. Increasing the variances for certain states (or models) will tend to decrease the distances between the input speech and those states. If this is done for model states corresponding to steady state regions, then the result will be to reduce the per frame contribution of discrepancies in those regions during recognition.

As described in Section 3 each state in a VFR model includes the mean of a feature which reflects the average number of frames in the original analysis condensed to a single frame corresponding to that state. These mean counts give some indication of how prolonged the corresponding sounds tend to be. It is therefore possible to weight each variance in a non-VFR model by multiplying by the corresponding mean count, n, from the equivalent VFR model.

Experiments were conducted on speaker RKM using the mean counts from a model file created from data with a VFR threshold of 350 to modify the variances of a model file created using the original (full-rate) data. Initially, all the three state models were modified, vowels and consonants. However, most of the frame reduction achieved by VFR corresponded to the centre states in the triphone models, therefore only the centre states were modified. It appeared to be possible to distinguish the vowels and consonants by the magnitude of these count mean values for state 2. Hence, experiments were conducted where the variances were only modified if the value of the mean count for state 2, n, was greater than some value. Also, rather then using n as the variance multiplication factor, various fractions of it were used since the effect on spectral distances is not strictly linear. These results are shown in Table 3.

The results in Table 3 are significantly worse than the word accuracy of 85.4% obtained for speaker RKM with a VFR threshold of 350. The words correct are all very similar. From this it would appear that VFR cannot be explained purely in terms of downweighting the steady state regions.

State 2 Mean	Scale Factor	Word correct	Word accuracy
unmodified		92.0%	79.8%
$n > 0$	n	91.7%	78.1%
$n > 2$	n	91.9%	79.1%
$n > 2$	$n/2$	92.2%	80.2%
$n > 2$	$n/4$	91.9%	78.7%
$n > 3$	n	92.0%	79.4%
$n > 3$	$n/2$	92.0%	79.6%
$n > 4$	n	92.0%	80.0%
$n > 4$	$n/2$	92.0%	80.0%

Table 3: Recognition results for speaker RKM and using the mean count for state 2 to modify the variances as shown.

7 Conclusions

It has been shown that VFR analysis can be successfully used within the ARM system without compromising performance levels.

It has been shown that VFR analysis can be used with triphone models with a consistent (though not necessarily significant) gain in performance at low thresholds. Even at low thresholds the amount of processing time is significantly reduced.

8 Future Work

All these results were obtained using data analysed at 100 frames per second. However, [1] suggested that the data should be analysed at 200 frames per second. Future work will investigate the effect of higher initial frame rates on the recognition performance obtained after VFR analysis.

It is hoped to carry out more experiments to investigate the effect of the DC offset on the data, and hence on the models created.

References

[1] J S Bridle and M D Brown, "A Data-Adaptive Frame Rate Technique And Its Use In Automatic Speech Recognition", Proc. Institute of Acoustics Autumn Conf.,Bournemouth, 9-10 November, pp C2.1-C2.6, 1982.

[2] J S Bridle and M D Brown and R M Chamberlain, "A One-Pass Algorithm for Connected Word Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, 1982, pp899–902.

[3] M J Hunt, "Evaluating the Performance of Connected Word Speech Recognition Systems" Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, New York, 1988, pp457–460.

[4] J N Holmes, "The JSRU Channel Vocoder", IEE Proceedings, vol 127, Part F, number 1, February 1980, pp 53–60.

[5] J N Holmes, "Speech Synthesis and Recognition", Van Nostrand Reinhold (UK), 1988.

[6] K-F Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System", PhD Thesis, Carnegie Mellon University, 1988.

[7] R K Moore, Private Communication.

[8] J Picone, G R Doddington and D S Pallett "Phone-Mediated Word Alignment for Speech Recognition Evaluation", unpublished draft ms., dated September 30, 1988.

[9] K M Ponting and S M Peeling, "Experiments in Variable Frame Rate Analysis for Speech Recognition", RSRE Memo 4330, 1989.

[10] K M Ponting and M J Russell, "The ARM Project: Automatic Recognition of Spoken Airborne Reconnaissance Reports", to appear in Proceedings of 'Military and Government Speech Tech 89', Arlington VA, 13- 15 November, 1989.

[11] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle, R K Moore,I Galiano and P Howell, "The ARM Continuous Speech Recognition System" to appear in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Albuquerque, New Mexico, 1990.

[12] D B Paul, "The Lincoln Robust Continuous Speech Recogniser", ICASSP 89, Glasgow, Scotland, May 1989.

REPORT DOCUMENTATION PAGE

DRIC Reference Number (If known) .....

Overall security classification of sheet .....Unclassified.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the field concerned must be marked to indicate the classification eg (R), (C) or (S).

Originators Reference/Report No.	Month	Year
MEMO 4336	FEBRUARY	1990

Originators Name and Location
RSRE, St Andrews Road
Malvern, Worcs WR14 3PS

Monitoring Agency Name and Location

Title
FURTHER EXPERIMENTS IN VARIABLE FRAME RATE ANALYSIS FOR SPEECH RECOGNITION

Report Security Classification	Title Classification (U, R, C or S)
Unclassified	U

Foreign Language Title (in the case of translations)

Conference Details

Agency Reference	Contract Number and Period

Project Number	Other References

Authors	Pagination and Ref
PEELING, S M; PONTING, K M	13

Abstract

The application of a simple variable frame rate analysis to the RSRE Airborne Reconnaissance Mission system, a continuous speech recognition system based on phone-level hidden Markov models, is described. Results are presented which show that performance using the variable frame rate technique and triphone models can be better than that obtained using triphone models and full frame rate data. The variable frame rate technique requires considerably less processing time.

Abstract Classification (U,R,C or S)
U

Descriptors

Distribution Statement (Enter any limitations on the distribution of the document)

Unlimited

ssoe